Rating scale design: a comparative study of two analytic rating scales in a task-based test

Bart Deygers, Koen Van Gorp, Lucia Luyten and Sien Joos

Katholieke Universiteit Leuven

Abstract

The Certificate of Dutch as a Foreign Language (Dutch: Certificaat Nederlands als Vreemde Taal, CNaVT), develops a suite of five domain-related task-based language tests, one of which is the test of Dutch for Academic Purposes. To ensure the representativeness of the task and the rating scale, the CNaVT is in the process of reshaping its rating scales.

The aim of this study was to develop and validate a new polytomous rating scale which incorporates both the CEFR and the opinions of subject specialists. In the research design both the existing dichotomous and the newly developed polytomous scales were used by four novice raters who each rated 125 written and spoken performances on the test of Dutch for academic purposes. The data emerging from the rating process was analyzed both qualitatively (semi-structured focus groups were conducted with the raters after the rating process) and quantitatively. Whereas the quantitative analysis indicates that the dichotomous scale is more reliable, the qualitative data offers a slightly different perspective, favouring the polytomous scale.

Rating scale development: general issues

The reliability and validity of different rating scale types has been the focus of linguistic research since the 1980s (see Barkaoui 2010 for an overview). In a more recent study Lumley (2002: 268) stresses the 'somewhat limited validity' of rating scales because of their 'inability to describe texts adequately'. The idea that any formalised description of language will be

unable to fully grasp all the subtleties of a real-life performance is shared by Fulcher (2010), who subdivides rating scales according to their construction process. He distinguishes measurement-driven rating scales from performance-driven scales. Measurement-driven scales are descriptions of linguistic performance that have been composed by language experts. Typically, measurement-driven rating scales are not based on real-life performances and since their abstract level descriptors may be distant from actual performance, these rating scales may lose any direct relationship to real performance. According to Fulcher 'This [measurement-driven] approach to scale design is primarily identified with the creation of the Common European Framework of Reference […]. Although the scale is empirically derived, it is not based on performance data, as there is no reference to the performance of learners or test takers on specific tasks, or even perceptions of the value of performances.' (Fulcher 2010: 7). Performance-driven rating scales on the other hand are constructed by closely analyzing and describing real-life performances, but may suffer from descriptional complexity (Fulcher 2010).

Rating scales can be subdivided according to the construction process, but categorisation can also be led by the way a scale yields a score, which can be holistic or analytic. In the former approach, raters judge a performance as a whole, whereas the latter compels raters to take into account separate features of language, such as grammar, vocabulary and structure (Alderson, Clapham and Wall 1995). In previous studies analytic scales have often proven to be more reliable than holistic ones (Weigle 2002, Knoch 2009, Barkaoui and Knouzi 2011) and to offer richer diagnostic information for L2 learners. Holistic scales, on the other hand have shown to be more authentic and quicker to use than their analytic counterparts (Weigle 2002, Knoch 2009).

To date, the effects of employing a holistic or an analytic rating scale have been researched with mixed results (Barkaoui 2010), but ultimately, it is not the rating scale but the

user of the scale, the rater, who has the final say (Lumley 2002). Naturally the descriptors' complexity and their level of abstraction will influence the quality of the judgments (Alderson et al 1995, Fulcher 2010), but it is the rater, who interprets the wording in the scale and who must maintain consistency of interpretation throughout the rating process. In order to enhance the reliability of the judgment, rater training is an effective tool (Shohamy, Gordon and Kraemer 1992, Weigle 1994, Lumley 2002) as it can influence future raters' behaviour and their interpretation of the criteria.

Combining a concern for rating scale type and rater expertise, Barkaoui (2010) studied the effect of using one type of scale together with either novice or experienced raters. Because of the way they are organised and subdivided, analytic scales were found to result in less conflicting decisions than holistic scales. Especially when performances are judged by novice raters, analytic scales are more reliable than holistic ones, since they increase self-consistency and focus the attention on the criteria at hand. Barkaoui notes that the type of rating scale steers the rater's judgement, implying that the type of scale used directly influences the validity and reliability of the test.

Both rating scales in the study at hand force the rater to make a number of judgments based on a series of criteria, both rating scales are analytic rather than holistic. In the dichotomous scale, which is currently used by the Certificate of Dutch as a Foreign Language (CNaVT), the rater is asked to judge a performance according to a series of criteria that are scored in a binary way. Polytomous scales on the other hand are not binary and share their architecture with band descriptors, similar to those found in the CEFR (2001) scales. As such, dichotomous rating scales force the rater to make a series of pass/fail decisions, whereas polytomous scales allow for more scoring options.

Context and general aims of the study

Since language testing is a practice that makes claims about real-life language ability based on performances which have often been gathered in artificial settings, the Certificate of Dutch as a Foreign Language (CNaVT) has purposefully moved away from indirect testing towards a direct, task-based approach (Ellis 2003, Gysen and Van Avermaet 2005). In order to minimise the gap between test performance and real-life ability, the CNaVT assesses whether a candidate can perform a real-life task to criterion. As such, it has adopted the Task-Based Language Assessment (TBLA) paradigm (Van Gorp and Deygers, in print) and it links in with Bachman's can-do typology of language tests (Bachman 2011).

The CNaVT offers a suite of five profile-based tests which fit three domains: societal, professional and educational. The distinction between a profile-based test and a domain-based test, is the scope and specificity of its context. Within one domain (i.e. the academic domain), the CNaVT distinguishes two profiles (i.e. Dutch for Academic Purposes - Students and Dutch for Academic Purposes - Teachers). The domains and profiles of the CNaVT were established in 2000 after an extensive needs analysis (Van Avermaet and Gysen 2006), based on the principles of Long (2005). This needs analysis determined to what end and in which contexts the CNaVT target audience uses Dutch. The profiles that emerged from the needs analysis encompassed a number of tasks types that could be considered representative within a specific domain.

One of the tests in the CNaVT's suite, the test of Dutch for Academic Purposes (Students), determines a candidate's ability to use and adapt language according to situational requirements (Davies 2001: 143) within an academic context. A college student, for example, should be able to send formal and informal e-mails, participate meaningfully in class, be a skilled writer and so on. Similar to the other profiles, the test of Dutch for Academic Purposes has been constructed with target language use, not language level as first priority. The test was therefore not specifically constructed for a given language level. Rather, it addresses

tasks that are considered representative for the target language use. After test construction, an expert panel linked the test of Dutch for Academic Purposes - Students to the B2 level of the Common European Framework of Reference for languages (CEFR). Since the 2008 CEFR alignment, the academic profile has been routinely checked and kept up to date, both in terms of content and in terms of cut-off score, which is monitored by means of a Rasch analysis after each test administration.

The pool of raters employed by the CNaVT may change from one test administration to the next, so novice raters are a common occurrence. Since novice raters appear to judge more reliably when using analytic scales and since analytic scales offer richer feedback data for second language (L2) learners, the CNaVT to date uses a dichotomous analytic rating scale. This scale gives the rater a series of criteria to score either 1 or 0. The pass/fail-logic shows that the rating scale was not designed to identify various language levels within one test but to make the distinction between test takers that are able to function in the target setting and those that are not.

When administering authentic task-based language tests, subject specialists may assist in determining task types and refining tasks that have been developed (Douglas 2000, 2001). In 2009, the subject specialists of the test of Dutch for Academic Purposes addressed an issue concerning the validity of the profile's dichotomous rating scale. According to them, the rating scales occasionally caused performances to be judged differently in the test than they would in real life. More concretely, the subject specialists assumed the scales might induce rater leniency when judging formal aspects of language.

This concern, raised by the subject specialists, instigated a redevelopment of the existing dichotomous rating scale as well as the development of a new polytomous scale. This new scale was to address the subject specialists' concerns, but also those of the end users, who whished for the rating scale to be linked to the CEFR more transparently. Following the

redevelopment, a study was conducted to compare the reliability and the validity of the polytomous and dichotomous scales employed.

Research Questions

The primary reason for conducting the study was to scrutinize the dichotomous rating scale of the test of Dutch For Academic Purposes (Students) and compare it to an alternative – polytomous – scale that addressed the needs and concerns raised by the CNaVT's subject specialists. These specialists had stressed the need for a rating scale that took into account their 'tacitly known criteria' (Jacoby and McNamara 1999: 224) and operationalised formal aspects of academic language. Additionally, the end users of the test of Dutch for Academic purposes had requested a clearer link between the rating scales and the CEFR.

But before the rating scales could be compared, a new rating scale had to be composed. This new rating scale was to meet the needs of the subject specialists and was to consider the relevant CEFR descriptors as reference points. Depending on the task, different CEFR scales were taken into account. Most widely used were the scales for Productive activities and strategies (pp. 57-65), Receptive activities and strategies (pp. 65-72), Interactive activities and strategies (pp. 73-84) and Linguistic competences (pp. 108-118). The challenge here was to create valid and reliable rating scales that did not suffer from the "descriptional inadequacy" sometimes associated with measurement-driven scales (Fulcher et al 2011).

The second goal of the study is research-based and includes two research questions: when compared both qualitatively and quantitatively

1. Is the new polytomous rating scale as reliable as the existing dichotomous scale?,
2. Is the new polytomous rating scale as valid as the existing dichotomous scale?

Development of the CNaVT scale

Subject specialist involvement

In the present study the subject specialists' involvement in the rating scale development process consists of two phases. In the first phase, subject specialists participated in an online questionnaire and in focus groups so as to generate rating criteria. In a second phase, the subject specialists offered feedback on the draft of the rating scale.

In order to determine the criteria to be used in a rating scale for a test of Dutch for Academic Purposes, two focus groups of domain experts were held, and an online questionnaire was administered to domain experts. The first focus group consisted of seven respondents, the second of six. Each focus group was attended by professionals employed within the academic target domain. All of them were regularly involved with student instruction and assessment of performances. Table 1 shows the professional background of the participants.

Table 1: Participants' professional background

Position	N
Language tutor in preparatory classes of Dutch for Academic Purposes	4
Academic staff (languages)	2
Academic staff (other subjects than languages)	3
Researcher	4

First, the focus group participants were asked to scrutinise the tasks of the test of Dutch for Academic purposes. By doing so, they got a thorough grasp of the task content, which

allowed them to make more meaningful comments concerning the tasks' rating scales. After this, the respondents were asked to voice the criteria they would employ when judging a performance on a note-taking task (audio input), a summarizing task (written input) and an argumentative speaking task (visual schematised input). The purpose of this was to tap into the subject specialists' real-life indigenous criteria (Jacoby and McNamara 1999). These are criteria that are often intuitively but not always consciously known by a group of people functioning within a specific field. A group of doctors for example might agree that a certain academic presentation is insufficient, although they may not immediately know why this is the case.

Having identified their criteria based on the task only, the subject specialists then received student performances on the same tasks. Now, the subject specialists were invited to refine or adjust their criteria. The criteria, which had been heavily content-focused after the first "blind" run, now became more focused on form.

Since each focus group served as a check of the other, the results from both focus groups were compared. Both focus groups showed an identical trend towards focus on form. A third source of information concerning intuitive rating criteria by academic staff was an online questionnaire comprising 178 subject specialists (see Table 2 for distribution according to profession).

Table 2: Distribution of questionnaire respondents according to profession

Position	N
Language tutor in preparatory classes of Dutch for Academic Purposes	34
Academic teaching position (languages)	41

Academic teaching position (other subjects than languages)	50
Researcher	57
Other	6

Offering similar questions as the ones asked in the focus groups, the purpose of the questionnaire was to check the generalisability of the conclusions drawn from the focus group. The data of the questionnaire largely replicates the criteria that had been identified in the focus group. The tendency to focus as heavily on form as on content was sustained. Table 3 lists the criteria that had been identified by the respondents for the three task types.

Table 3: Respondents' criteria for three task types

Note taking task	Summarising task	Argumentative task
Skill: Integrated writing	Skill: Integrated writing	Skill: Integrated speaking
Expected performance: notes	Expected performance: written summary	Expected performance: argumentative speaking
Audio input	Written input	Schematised input

Criteria	%	Criteria	%	Criteria	%
Content (accuracy)	27,8	Structure	16,9	Structure	18,5
Grammar	11,3	Content (accuracy)	15	Content (accuracy)	16,1
Structure	10,7	Summarizing skills	13,4	Grammar	10,6

Spelling	4,8	Grammar	13,1	Argumentation	9,6
Vocabulary	4,4	Style	6,7	Pronunciation and fluency	7,5

After a first draft rating scale was composed based on the criteria that had been identified by the focus group members and by the respondents of the questionnaire, the focus groups were invited again to offer feedback on the redesigned rating scale. Because subject specialists were consulted on several occasions throughout its composition, the polytomous scale went through an iterative and dialogic development process.

The subject specialist input led to a number of important changes in the rating scale design, the most important one being the increase of the relative importance of formal aspects of language. The polytomous scale takes into account similar formal criteria as the dichotomous scale but their relative weight in comparison with content is greater than in the dichotomous scale. As such, the polytomous scale focuses on getting the message across appropriately in an academic context.

The CEFR and the Development of the polytomous scale

Whereas the original dichotomous rating scale presents two options for each criterion, the polytomous scale offers four, ranging from unsatisfactory to excellent. The B2-target level occupies the third level in the scale (Table 4).

The criteria for each task were based on the parameters that were identified by the subject specialists who attended the focus groups and those who filled out the online questionnaire. For the argumentative speaking task, the following criteria were operationalised: register (a criterion that was stressed by the focus groups members, but not by the respondents of the

questionnaire), content, argumentation, structure & cohesion, vocabulary, grammar and pronunciation and fluency. The wording for each level was adapted from the CEFR according to the requirements of the task. Even though the CEFR 'was not designed specifically for test specifications and language testing contexts' (North 2004 in Papageorgiou 2010: 273), it remained a crucial point of reference during the rating scale composition process.

Table 4: Polytomous rating scale layout

Target level +1	C1
Target level	B2
Target level -1	B1
Target level -2	A2

After the scale had been composed and approved by the subject specialists, it was trialled in two pilot studies (see below). First, four raters used the scales to rate a total of 250 performances. After this, they were invited to offer feedback on the usability and interpretability of the scale. The raters reported vagueness of the level descriptors as the main problem when using the polytomous scale. Based on the feedback, the scales were reformulated and the scales were piloted again with a second team of novice raters. During rater training, the raters were given the chance to think about alternative wording to make the scales more easily interpretable.

The rewritings focused on simplifying the sometimes overly abstract CEFR-based descriptors and on marking the borders between levels more clearly. Vagueness was avoided as much as possible by replacing such terms as 'adequate' and 'nearly perfect' with more readily interpretable alternatives that can be grouped into four categories: concrete insertions, subjective insertions, discriminating insertions and exemplary additions.

Concrete insertions are additions to the CEFR descriptors that serve to give the rater a better foothold. One such insertion in the criterion argumentation is 'the argumentation is unconvincing and cannot be maintained without the interlocutor's help' (target level-2). A subjective insertion takes on the perspective of the novice rater when listening to concrete performances: 'the structure is consistent and perfectly aligned with the content. The audience has no problem following the presentation' (Target level) / 'Every now and then, the audience may lose track of the presentation' (Target level-1). Discriminating additions serve to make the borders between two performance levels more clear. In one instance, the wording changed from 'The performance is largely understandable' to 'The performance is only partly understandable' (Target level-1). Exemplary additions are concrete grammatical markers of ability. They are syntactic structures that should be mastered at a given level and increase rater confidence because they are concretely observable: 'in complex structures grammatical flaws may occur even though common grammatical structures (e.g. conjugation, inversion, and subclause) are mostly correct.' / 'The performance shows mastery of basic grammatical patterns (simple clauses, main word order)'.

By the time of its completion, the polytomous rating scale had received input from the CEFR, from subject specialists, from raters and from test developers. This resulted in a four-point scale which focuses on getting the message across adequately in an academic context. The example below shows the descriptors for rating "Structure and Cohesion" in a presentation task in the dichotomous scale (Table 5) and the polytomous scale (Table 6).

Table 5: Dichotomous scale for "Structure and Cohesion"

	1	0
The text is well structured. It has a clear organization and uses cohesive devices		

Table 6: Polytomous scale for "Structure and Cohesion"

Structure & cohesion	
The presentation's structure is consistent and perfectly aligned with the content. The audience has no problem following the presentation. The presentation shows a varied and correct use of cohesive devices and structuring strategies. The presentation's structure supports its content.	A
The presentation is coherent and for the most part logically structured. Every now and then, the audience may lose track of the presentation. *Cohesive devices are mostly used correctly. Largely coherent even though some parts of the communication are not always effective.*	B
The presentation shows "jumpiness" and/or is occasionally lacking in cohesion. *Cohesive devices are limited to common linear linking words. Due to lack or misuse of cohesive devices, the internal cohesion may be insufficient.*	C
The presentation hardly shows structure or cohesion. *Fragmentary or inadequately structured to such an extent that the intended message is difficult to understand.*	D

Piloting of the polytomous CNaVT scale

Research design

As discussed above, the rating scale research served to gather information on the usability of the scales. Additionally, the research compared the dichotomous scale to the polytomous scale in terms of reliability and validity.

The quantitative data collecting process involved four novice raters who were paired. For both rating scales, the raters received a two-day training, which has been shown to positively influence rater reliability (Knoch 2009). After the training, the raters received a *decision booklet*, which enlists a number of rating pitfalls and shows how to go about them as a rater (Knoch 2009).

Two by two the raters rated the same task performances (N = 250). The first pair of raters rated 125 task performances (75 integrated writing tasks and 50 integrated speaking tasks, clustered in 25 tests) first with the polytomous scale and subsequently with the dichotomous one. In order to monitor the influence of one rating scale on the next, the second dyad took the reverse order and started out by rating their 125 performances with the dichotomous scale first and the polytomous one next. All performances used in this research had been preselected so as to guarantee varying levels of linguistic ability, geographical dispersion and a wide array of first languages (L1s).

Since the raters ware paired throughout the rating process and each pair judged the same 125 performances twice, the main variable was the rating scale used. Each scale was used to rate a total of 250 task performances and was used by two pairs of raters in two different sequences. The data gathered is the result of four different setups: rater A/B Polytomous, rater A/B dichotomous, rater C/D dichotomous and rater C/D polytomous (Table 7).

Table 7: Quantitative research design

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous ↓ Dichotomous	Dichotomous ↓ Polytomous

For each different setup, the same data analysis occurred, i.e. a Pearson correlation to determine the strength of the connection between the ratings of each rater dyad. Secondly, Cohen's Kappa was calculated, so as to illustrate the measure of agreement between the raters (North 2009).

After the rating procedure was concluded, the qualitative data was gathered. The first four raters were invited to a semi-structured focus group to discuss the usability and interpretability of the rating scales and the level descriptors. Based on their comments the rating scales were adjusted and a second group of five raters was called upon. During their training, this second group of raters helped to adjust level descriptors that were considered multi-interpretable. After this, they rated additional performances (N=76) and took part in a new semi-structured focus group which served to determine whether the changes to the level descriptors had increased the rating scale's usability.

Findings

Quantitative findings

Rater A and B rated their 125 tasks with the polytomous scale first, yielding a medium positive relationship ($r = .47$, $p < .01$), a fair inter-rater agreement ($K = .030$) and an α of .76. When considering the data closely, little correspondence between the judgments can be found around the cut-off score. Rater A agrees with 40 of the 65 performances that rater B, considers adequate. Of the 25 remaining performances, rater A considers 10 to be insufficient and 15 to be excellent. There is no real consistency to be found in the cases of non-agreement. The unsatisfactory correlation and the low measure of rater agreement may indicate that the polytomous scale allows for multi-interpretability.

Next, rater A and B rerated the 125 tasks using the dichotomous model. Here, the correlation shows a strong positive relationship ($r = .82$, $p < .001$), the rater agreement is moderate ($K = .59$) and the reliability has increased slightly ($\alpha = .77$). Around the cut-off, there is more agreement between the raters than there was when using the polytomous scale. When closely examining the 90 performances that rater A considers sufficient, there are 25 performances considered less than adequate by rater B. Rater B's severity is consistent

however, so the pattern around the cut-off score is less random than it was when using the polytomous scale.

So as to get a grip on the effect of the order in which the scales were used, rater C and D took the reverse order from rater A and B and began rating their performances with the dichotomous scale first and the polytomous one next. The ratings of Rater C and D, rating different performances than rater A and B, also showed a strong positive relationship when using the dichotomous scale ($r = .94$, $p < .001$) as well as an increased reliability ($\alpha = .86$). The inter-rater agreement was moderate ($K = .54$), but after removing 13 items out of 56 items with a negative Item-Total Correlation (no negative ITCs were observed for rater A and B when using the polytomous scale), the inter-rater agreement was perfect ($K = 1$).

The data emerging from the polytomous rating process of rater C and D shows a lower correlation than was the case when using the dichotomous scale ($r = .79$, $p < .001$). The measure of rater agreement is fair ($K = .35$) and the reliability remained unchanged ($\alpha = .86$). When examining the items clustered around the cut off score however, little to no agreement is to be found.

Qualitative findings

Contrary to what Knoch (2008) found in her study, he raters who took part in this study showed no tendency to neglect the outer bands of the polytomous scale. During the focus group after the rating process, they did report other effects from using a polytomous scale. One such effect was the subjective judgment of time spent rating exams. The four novice raters involved in the focus group reported having spent at least 50% more time when using the polytomous scale. In reality, using polytomous scales to rate performances indeed took longer, but the difference was roughly 20%. All respondents claimed that the polytomous scale required them to consider the performance and the descriptors more thoroughly,

required them to reflect more. The continual conscious reflection could be what caused the respondents to believe the polytomous scale to be disproportionately time-consuming. If anything, this implies that the respondents did not discard the outer bands of the scale but rather took them into account for every performance.

When asked about their preferred scale in general and for their preferred scale to assess written performances, all raters opted for the dichotomous one, stating that it appeared to be faster than the polytomous one. Additionally, they appreciated the sense of certainty the dichotomous options induced, as opposed to the confusion sometimes induced by a polytomous scale. The limited number of available options in the dichotomous rating scale makes this type of scale easier to memorise than its multi-faceted counterpart, but according to the respondents it is also what makes it too crude an instrument. Three out of four novice raters taking part in the current study considered the polytomous scale ideal for assessing speaking tasks. According to the raters, the polytomous scale's advantages are also its downsides: because it is in line with one's intuitive judgment, it may lead to subjective rating. The raters feared that their experience with rating was too limited to allow for intuition to influence their judgement. For that reason, they preferred more guidance Also, since the polytomous scale allows for more detailed judgment, the raters reported, it may cause doubt. In short, the novice raters involved in this pilot found the polytomous rating scale too vague.

Since the raters reported the occasional vagueness of the level descriptors in the polytomous scale as the cause of the doubt when rating, the descriptors for oral production were rewritten (cf. above). These rewritten descriptors were piloted in a small-scale trial which served to check whether it would be possible to improve the strong aspects of the polytomous scale while diminishing its negative effects. The trial involved five novice raters who rated a total of 76 oral performances using rewritten level descriptors.

After rating 76 oral production tasks, four out of five raters partaking in the semi-structured focus group reported preferring the polytomous scale for oral production to the dichotomous one. The raters who preferred the polytomous model did so because they no longer believed the rating scale to be too vague or too abstract. Additionally, they reported it to allow for fine-grained distinctions between language levels.

Discussion

The extent of the difference in terms of reliability (α) and rater agreement (K) between both scales can clearly be observed in Table 10, which shows that the rater agreement is consistently lower when using the polytomous scale. Raters A and B, who started off with the polytomous scale show less agreement on the polytomous scale than rater C and D, who may have benefited from the sequence effect. Still, irrespective of the order in which the performances were rated, the dichotomous scale emerges as the most reliable option. This implies that even if the sequence of rating would have had an impact on the reliability of the rating, it would not necessarily have benefited the dichotomous scale. Indeed, the reliability indexes of raters C and D are highest for the dichotomous scale, which was used first. So, even if the order in which the scales were used would have had an effect, the dichotomous scale used consistently outperforms the polytomous one in terms of rater correspondence and inter-rater agreement.

Table 12: Reliability indexes of polytomous and dichotomous scales

Rater A/B Performance 1-125	Rater C/D Performance 126-250
Polytomous	Dichotomous
r = .47	r = .94
K = .30	K = .54
α = .76	α = .86
↓	↓

Dichotomous	Polytomous
r = .82	r = .79
K = .59	K = .35
α = .77	α =.86

The quantitative differences between both scales are to some extent mathematically explainable. Indeed, it is normal for correlations to be more robust as the number of options decreases. Likewise, it is a known fact that 'unweighted kappa coefficients decrease with the number of categories' (Brenner and Kliebst 1996: 199) and that impressionistic descriptors provide 'a wider window for rater interpretation of the meaning of the descriptors, but […] inevitably results in lower inter-rater reliability' (Knoch 2008: 61). Therefore, even though the differences in reliability indices for both rating scales should not be ignored, the quantitative data should be supplemented with qualitative input, which offers information on the validity and interpretability of the scales.

In the focus groups conducted with the first team of novice raters, they reported a sense of certainty caused by dichotomous options and the confusion the CEFR-based polytomous descriptors sometimes caused. This may help to explain the quantitative differences between both scales. Additionally, the qualitative follow-up study shows that through the process of actively exploring and refining the rating scale together with prospective raters affects rating behaviour. The rewritings focused on simplifying the sometimes overly abstract CEFR-based descriptors and on marking the borders between levels more clearly. Vagueness was avoided as much as possible by using the concrete insertions, subjective insertions, discriminating insertions and exemplary additions discussed above.

Conclusion

This study compared a polytomous to a dichotomous analytic rating scale in terms of reliability and validity. The polytomous rating scale was developed in close conjunction with subject specialists and consists of CEFR-like level descriptors, whereas the dichotomous (also the scale which is currently in use) is made up of a series of binary options. Even though the dichotomous scale includes descriptors concerning the formal aspects of language, the focus is on the content, on getting the message across. The polytomous scale focuses on getting the message across appropriately, thereby giving a larger proportional weight to formal aspects of language, such as structure, register and grammatical accuracy.

Both rating scales were piloted with two pairs of novice raters. Each pair rated 125 performances, the first pair starting off with the polytomous scale and switching to the dichotomous one, the other pair using the reverse order. The results from the quantitative data show the dichotomous scale to be consistently more reliable, irrespective of the order in which the raters used the scales.

Overall, the raters preferred the dichotomous scale to the polytomous one, because having two instead of four categories made them feel more certain about their decision-making process, but also because it was less intuitive and less vague. The vagueness of the polytomous descriptors most likely stemmed from the fact that they had been composed together with subject specialists as well as testing specialists and that they had been based on CEFR descriptors, which may appear too vague for novice raters, even after rater training. The fact that the polytomous scales were considered to leave too much room for interpretation and that these scales proved to be less reliable than their dichotomous counterparts, links in with Bachman's observation that 'vagueness in task specification inevitably leads to vagueness in measurement' (Bachman 2002: 458).

The raters' preference for the dichotomous scale did contain one important exception, since they preferred the polytomous scale for assessing speaking. In a follow-up study, five new novice raters were called upon to assess speaking tasks by means of the polytomous scale which had been adjusted in line with the results from the first pilot. During the rater training for the second pilot, the raters were invited to comment on the scale descriptors in order to help reformulate them by using words they could grasp more easily. The wording of the new scales consequently moved away from the CEFR terminology and became more tangible for novice raters. The process of thinking about the rating scale and making it 'more detailed, empirically-developed' did result 'in subsequent changed rating behaviour' (Knoch 2008: 62). The raters involved in this second pilot (containing only oral production tasks) largely preferred the polytomous scale - which they had helped rewrite - to the dichotomous one. Future quantitative analyses will be necessary to investigate whether this effect has also improved the reliability of the rating scale.

Even though this study has reaffirmed the statistic robustness of the dichotomous scale, the subject specialists and the raters involved in both pilot studies indicated its limitations regarding authenticity and validity. After revising the polytomous scale together with the end users, its interpretability had improved.

Further quantitative research will be needed to determine the reliability of the rewritten polytomous rating scales. If the reliability indexes in a new large-scale pilot are satisfactory, the scales for written production will be rewritten parallel to those for speaking.

References

Alderson, C J, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Bachman, L F (2002) Some reflections on task-based language performance assessment *Language Testing*, 19, 454–76.

Bachman, L F (2011, July) How Do Different Language Frameworks Impact Language Assessment Practice, Plenary talk at ALTE 4th International Conference, Kraków, Poland.

Bachman, L and Palmer, A (2010, July) *Language Assessment in Practice*, Oxford: Oxford University Press.

Barkaoui, K (2010) Explaining ESL essay holistic scores: A multilevel modeling approach, *Language Testing*, 27, 515-535

Barkaoui, K and Knouzi, I (2011, July) Rating scales as frameworks for assessing L2 writing: examining their impact on rater performance, Paper presented at ALTE 4th International Conference, Kraków, Poland.

Brenner, H and Kliebst (1996) Dependence of weighted kappa coefficients on the number of catergories, *Epidemiology*, 7, 199-202.

Cohen, J (1960) A coefficient for agreement for nominal scales, *Education and Psychological Measurement*, 20, 37–46.

Colpin, M and Gysen, S (2006) Developing and introducing task-based language tests, in Van den Branden, K (Ed), *Task-based language education: from theory to practice*, Cambridge: Cambridge University Press, 151–74.

Davies, A (2001) The logic of testing languages for specific purposes, *Language Testing*, 18, 133–47.

Douglas, D (2000) *Assessing languages for specific purposes*, Cambridge: Cambridge University Press.

Douglas, D (2001) Language for specific purposes assessment criteria: where do they come from?, *Language Testing*, 18, 171–185.

Ellis, R (2003) *Task-based language learning and teaching*, Oxford: Oxford University Press.

Fulcher, G, Davidson, F and Kemp, J (2011) Effective rating scale development for speaking tests: Performance decision trees, *Language Testing*, 28, 5-29

Gysen, S, and Van Avermaet, P (2005) Issues in Functional Language Performance assessment: The Case of the Certificate Dutch as a Foreign Language, *Language Assessment Quarterly*, 2, 51–68.

Jacoby, S, and McNamara, T (1999) Locating Competence, *English for Specific Purposes*, 18, 213–41.

Knoch, U (2008) The assessment of academic style in EAP writing: The case of the rating scale, *Melbourne Papers in Language Testing,* 13, 34-67.

Knoch, U (2009) Diagnostic assessment of writing: A comparison of two rating scales, *Language Testing*, 26, 275–304.

Long, M (2005) *Second language needs analysis*, Cambridge: Cambridge University Press.

Lumley, T (2002) Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19, 246–276.

McNamara, T and Roever, C (2006) *Language Testing: The Social Dimension*, Malden, MA: Blackwell Publishing.

McNamara, T (2006) Validity in Language Testing: The Challenge of Sam Messick's Legacy, *Language Assessment Quarterly*, 3, 31–51.

Norris, J M (2009) Task-Based Teaching and Testing, in Long, M H and Doughty C J (Eds) *The handbook of Language Teaching*, Malden, MA: Wiley-Blackwell, 578–94.

Norris, J M, Brown, J D, Hudson, T D, and Bonk, W (2002) Examinee abilities and task difficulty in task-based second language performance assessment, *Language Testing*, 19, 395–418.

North, B (2009) Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual, Language Policy Division, Strasbourg.

Papageorgiou, S (2010) Investigating the decision-making process of standard setting participants, *Language Testing*, 27, 261–82.

Sawaki, Y (2007) Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite, *Language Testing*, 24, 355–90.

Shohamy, E, Gordon, C and Kraemer, R (1992) The effect of raters' background and training on the reliability of direct writing tests, *Modern Language Journal*, 76, 27-33.

Shohamy, E (1996) Language testing: Matching assessment procedures with language knowledge, in Birenbaum, M and Dochy, F (Eds), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*, Boston, MA: Kluwer Academic Publishers, 143–59.

Skehan, P (1996) A framework for the implementation of task based instruction, *Applied Linguistics*, 17, 38–62.

Skehan, P (1998) *A cognitive approach to language learning*, Oxford: Oxford University Press.

Spence-Brown, R (2001) The eye of the beholder: authenticity in an embedded assessment task, *Language Testing*, 18, 463–81.

Van Avermaet, P, and Gysen, S (2006) From needs to tasks: Language learning needs in a task-based approach, in Van den Branden, K (Ed), *Task-based language education: from theory to practice*, Cambridge: Cambridge University Press, 17–46.

Van Gorp, K and Deygers, B (in print). Task Based Language Assessment, in Kunnan, A (Ed.) *The Companion to Language Assessment*. New Jersey: Wiley-Blackwell.

Van den Branden, K (Ed) (2006) *Task-based language education*, Cambridge: Cambridge University Press.

Weigle, S C (1994) Effects of training on raters of ESL compositions, *Language Testing*, 11, 197-223.

Weigle, S C (2002) *Assessing writing*, Cambridge: Cambridge University Press.

Wigglesworth, G (2008) Task and Performance Based Assessment, in Shohamy, E and Hornberger, N H (Eds), *Encyclopedia of Language and Education*, New York: Springer, 111–122.

Wu, W M and Stansfield, C W (2001) Towards authenticity of task in test development, *Language Testing*, 18, 187–206.